(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **2002189492 A**

(43) Date of publication of application: **05.07.02**

(51) Int. Cl **G10L 15/06**
G10L 15/10
G10L 15/14
G10L 15/20
G10L 21/02

(21) Application number: **2000385212**

(22) Date of filing: **19.12.00**

(71) Applicant: **SHARP CORP**

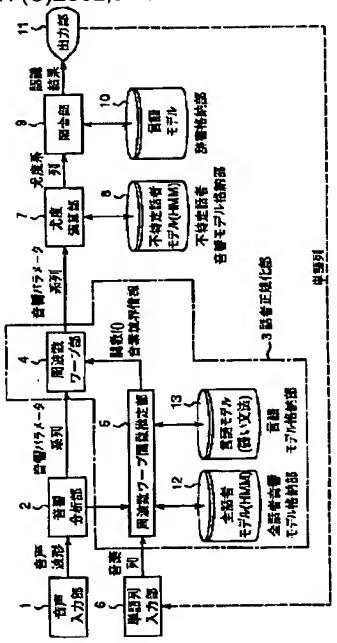
(72) Inventor: **YAMAGUCHI KOICHI**
HACHIMAN YOICHIRO

(54) APPARATUS AND METHOD FOR EXTRACTING SPEAKER'S FEATURE, SPEECH RECOGNITION DEVICE, SPEECH SYNTHESIZER, AND PROGRAM RECORDING MEDIUM

(57) Abstract:

PROBLEM TO BE SOLVED: To stably extract a speaker's feature without depending on the contents of utterance data.

SOLUTION: A frequency function estimation part 5 is provided with a phonemic limit estimation part, a frequency measuring part, and a mode extracting part. During learning, the frequency function estimation part 5 estimates phonemic limit information, and performs the maximum likelihood estimation of the coefficient α of the frequency warping function f of each sample about a phonemic section selected on the basis of the phonemic limit information. Furthermore, the frequency function estimation part 5 determines a distribution function H (α) which represents a frequency distribution about the coefficient αof each sample, and estimates a coefficient a which provides a mode value as the optimal coefficient of the frequency warping function f. Consequently, a correct frequency warping function f can be estimated even when a plurality of peaks are present in the frequency distribution, and the speaker's feature can stably be extracted without depending on the contents of utterance data.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－189492
（P2002－189492A）
(43)公開日　平成14年７月５日(2002.7.5)

(51)Int.Cl.[7]	識別記号	Ｆ Ｉ		テーマコード（参考）
Ｇ１０Ｌ　15/06		Ｇ１０Ｌ　3/00	５２１Ｓ	５Ｄ０１５
15/10			５３１Ｊ	
15/14			５３１Ｃ	
15/20			５３５Ｂ	
21/02		3/02	３０１Ａ	

審査請求　未請求　請求項の数10　ＯＬ　（全 19 頁）

(21)出願番号　特願2000－385212(P2000－385212)

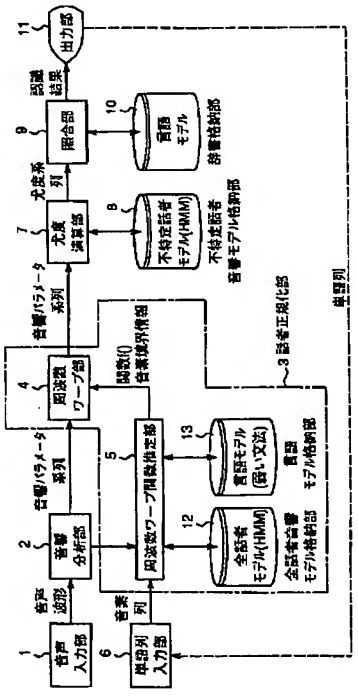
(22)出願日　平成12年12月19日(2000.12.19)

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号
(72)発明者　山口　耕市
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(72)発明者　八幡　洋一郎
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(74)代理人　100062144
弁理士　青山　葆　（外１名）

Ｆターム(参考)　5D015 AA02 BB02 EE03 HH23

(54)【発明の名称】　話者特徴抽出装置および話者特徴抽出方法、音声認識装置、音声合成装置、並びに、プログラム記録媒体

(57)【要約】
【課題】　発声データの内容に依存せず安定して話者特徴を抽出する。
【解決手段】　周波数関数推定部５には音素境界推定部，頻度計測部およびモード抽出部を備えている。そして、学習時には、音素境界情報を推定し、この音素境界情報に基づいて選択した音素区間に関して、各サンプル毎に、周波数ワーピング関数fの係数αを最尤推定する。さらに、各サンプル毎の係数αに関する頻度分布を表す分布関数Ｈ(α)を求め、最頻値を与える係数αを周波数ワーピング関数fの最適係数として推定する。したがって、上記頻度分布に複数のピークが存在する場合でも正確な周波数ワーピング関数ｆを推定でき、発声データの内容に依存せず安定して話者特徴を抽出できる。



1

【特許請求の範囲】

【請求項１】　入力話者の音声から、標準話者の音声スペクトルに対して上記入力話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を話者特徴として抽出する話者特徴抽出装置において、

所定の音声単位毎に、上記標準話者の音響モデルに対して、上記入力話者の音声サンプルの尤度あるいは音響モデルを上記入力話者の音声サンプルに話者適応させた話者適応音響モデルの尤度を最大にするという基準に従って、上記周波数伸縮関数を最尤推定し、この推定された上記周波数伸縮関数の集合の頻度分布を求める頻度計測手段と、

上記頻度分布に基づいて、最大頻度を有する周波数伸縮関数を話者特徴として抽出するモード抽出手段を備えたことを特徴とする話者特徴抽出装置。

【請求項２】　請求項１に記載の話者特徴抽出装置において、

上記入力話者の音声サンプルから音響モデルを用いたビタビアルゴリズムによって音素境界情報を推定する音素境界情報推定手段を備えて、

上記頻度計測手段は、上記音素境界情報に基づいて、有音無音の別および調音点の位置に従って、上記最尤推定を行う際に用いる上記入力話者の音声サンプルの音声区間を限定する機能を有していることを特徴とする話者特徴抽出装置。

【請求項３】　請求項１に記載の話者特徴抽出装置において、

音響モデルの各状態が、有音無音の別および調音点の位置に従って予め設定された上記最尤推定を行う音声区間に属しているか否かを判別し、属している状態に関して、上記音響モデルを入力音声サンプルに話者適応させて話者適応音響モデルを作成する話者適応モデル作成手段を備えて、

上記頻度計測手段は、上記標準話者音響モデルに対して上記話者適応音響モデルを用いて上記最尤推定を行うようになっていることを特徴とする話者特徴抽出装置。

【請求項４】　請求項１乃至請求項３の何れか一つに記載の話者特徴抽出装置において、

上記モード抽出手段は、上記最大頻度を有する周波数伸縮関数が複数存在する場合には、上記頻度分布を混合ガウス分布で表現した場合における当該複数の周波数伸縮関数が属している分布の分散が大きい方の周波数伸縮関数をもって話者特徴する機能を有していることを特徴とする話者特徴抽出装置。

【請求項５】　請求項４に記載の話者特徴抽出装置において、

上記モード抽出手段は、上記標準話者の特徴を表す周波数伸縮関数に近い方の周波数伸縮関数をもって話者特徴とする機能を組み合せて、上記話者特徴を抽出するようになっていることを特徴とする話者特徴抽出装置。

2

【請求項６】　標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を用いて入力話者の音声スペクトルの周波数軸を伸縮することによって上記入力話者の音声を正規化する正規化手段を有する音声認識装置において、

上記正規化手段は、

請求項１乃至請求項５の何れか一つに記載の話者特徴抽出装置と、

上記話者特徴抽出装置によって抽出された周波数伸縮関数を用いて、上記入力話者の音声スペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声認識装置。

【請求項７】　入力話者の音声スペクトルに対して標準話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を用いて音声のスペクトルの周波数軸を伸縮することによって音響モデルを入力話者に話者適応させる話者適応手段を有する音声認識装置において、

上記話者適応手段は、

請求項１乃至請求項５の何れか一つに記載の話者特徴抽出装置と、

上記話者特徴抽出装置によって抽出された周波数伸縮関数の逆関数を用いて、上記音響モデルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声認識装置。

【請求項８】　入力話者の音声スペクトルに対して標準話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を用いて音声のスペクトルの周波数軸を伸縮することによって、標準話者の音声素片を接続して成る合成音声の声質を発話者の声質に変換する声質変換手段を有する音声合成装置において、

上記声質変換手段は、

請求項１乃至請求項５の何れか一つに記載の話者特徴抽出装置と、

上記話者特徴抽出装置によって抽出された周波数伸縮関数の逆関数を用いて、上記音声素片の周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声合成装置。

【請求項９】　入力話者の音声から、標準話者の音声スペクトルに対して上記入力話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を話者特徴として抽出する話者特徴抽出方法において、

所定の音声単位毎に、上記標準話者の音響モデルに対して、上記入力話者の音声サンプルの尤度あるいは音響モデルを上記入力話者の音声サンプルに話者適応させた話者適応音響モデルの尤度を最大にするという基準に従って、上記周波数伸縮関数を最尤推定し、

この推定された上記周波数伸縮関数の集合の頻度分布を求め、

上記頻度分布に基づいて、最大頻度を有する周波数伸縮関数を話者特徴として抽出することを特徴とする話者特

徴抽出方法。

【請求項１０】　コンピュータを、

請求項１における上記頻度計測手段およびモード抽出手段として機能させる話者特徴抽出処理プログラムが記録されたことを特徴とするコンピュータ読出し可能なプログラム記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、標準話者の音声スペクトルに対する入力音声スペクトルの周波数軸の線形伸縮関数を話者特徴として抽出する話者特徴抽出装置および話者特徴抽出方法、その抽出方法を用いた音声認識装置,音声合成装置、並びに、話者特徴抽出処理プログラムを記録したプログラム記録媒体に関する。

【０００２】

【従来の技術】従来より、隠れマルコフモデル(Hidden Markov Model：以下、ＨＭＭと言う)を用いた音声認識方法の開発が近年盛んに行われている。このＨＭＭは、大量の音声データから得られる音声の統計的特徴を確率的にモデル化したものであり、このＨＭＭを用いた音声認識方法の詳細は、中川聖一著「確率モデルによる音声認識」(電子情報通信学会)に詳しい。このＨＭＭに基づく話者適応や話者正規化に関する研究が行われている。ところが、通常の話者正規化技術や話者適応技術においては発声データの内容や量に依存するので、少量の発声データからでは安定した性能向上が非常に難しい。そこで、声道長を用いた話者適応や話者正規化の手法が注目されており、特に声道長に基づく話者正規化が盛んに研究されて効果が出ている。声道長は音声のスペクトルの大まかな特徴を表すパラメータであり、声道長の差は話者間の主な変動要因である。また、声道長は従来の話者適応法に比べて１個のパラメータあるいは極めて少ないパラメータで音声の特徴を表現できることから、声道長にはより少量の学習データで効率良く正規化できるというメリットがある。

【０００３】ところで、標準話者の音声パターンに対する入力話者の音声サンプルの尤度を最大にするという基準(最尤推定)で、上記音声サンプルにおける周波数軸の線形伸縮係数α(声道長正規化係数)を求める方法(ML-VTLN法：Maximum LikelihoodVocal Tract Length Normalization)がある。そして、この声道長正規化係数αを用いて入力話者の音声サンプルの周波数軸を線形伸縮することで話者正規化を行う技術が提案されている(例えば、AT&T Bell Labs. Li Lee, Richard C.Rose, “Speaker Normalization using Efficient Frequency Warping Procedures”,pp.353-356 ICASSP96 (1996))。また、特開平１１－３２７５９２号公報においては、声道を前室と後室との２つの室に分け、入力音声のフォルマント周波数を用いて各室に対応した２つの周波数軸線形伸縮係数αを求め、この２つの周波数軸線形伸縮係数αを用いて話者正規化する技術が開示されている。

【０００４】尚、上記話者適応は標準となる音響モデルを入力話者に対して適応(つまり正規化)させる技術であり、話者正規化とは表裏一体の関係にある。

【０００５】さらに、音声合成における声質変換に関する従来技術として、音声認識の話者適応技術を用いてスペクトルの写像を行なう方法が提案されている。例えば、ベクトル量子化(ＶＱ)コードブックマッピング法をベースとした話者適応技術を用いる方法(特開平１－９７９７号公報)や、ＶＦＳ(Vector Field Smoothing)法をベースとした話者適応技術を用いる方法(橋本誠,樋口宣男:“話者選択と移動ベクトル場平滑化を用いた声質変換のためのスペクトル写像”,信学技報, SP95-1, p.p.1-8, May 1995)等がある。

【０００６】

【発明が解決しようとする課題】しかしながら、上記従来の声道長に基づく話者適応や話者正規化には、以下のような問題がある。すなわち、声道長に基づく話者適応や話者正規化は極めて少ないパラメータ数で音声の特徴を表現できるとは言うものの、声道長の抽出は発声データの内容や量に大きく左右され易い。したがって、必ずしも少ない学習サンプルから安定して声道長を抽出できるとは限らない。その結果、声道長に基づいて話者正規化や話者適応や話者クラスタリングを行うような音声認識装置においては、性能劣化を招くという問題がある。

【０００７】上記声道長正規化係数αを求める方法としては、上述したように、学習サンプル全体を対象として最尤推定するML-VTLN法が提案されている。この方法においては、発話の仕方や発話内容によってスペクトルの概形が変動するので、学習サンプルによって最適な声道長正規化係数αが異なってしまう場合が生ずる。つまり、異なる周波数軸伸縮関数で局所的に最適となるため、声道長正規化係数αの頻度分布に複数のピークが生じてしまうという多峰性の問題が発生し、真の声道長正規化関数が安定して推定できないことになる。

【０００８】また、声道長の伸縮(周波数軸のワーピング)を線形関数やそれに類似した関数で表現しており、一般に全音素区間に対して周波数ワープを学習・作用するようにしている。そのために、声道長正規化係数αを用いて話者正規化を行う方法においては、声道長の差の影響を受け難い音素や無音部まで学習および正規化してしまうという問題もある。

【０００９】すなわち、上記従来の声道長正規化係数の推定方法では、正確な声道長正規化係数が推定できなかったり、入力音声サンプルを必要以上に変形してしまったりするために、話者正規化に適用した場合には、認識性能の低下を招くことになるのである。

【００１０】さらに、上記特開平１１－３２７５９２号公報の正規化方法においては、２つのサンプルから直接声道パラメータを求めるようにしているが、声道パラメ

ータを得るために入力音声のフォルマント周波数を用いている。ところが、一般的にフォルマント周波数を全自動で求めることは困難であり、上記特開平１１－３２７５９２号公報に開示された線形伸縮係数を用いた話者正規化方法では、実時間性に欠けるという問題がある。

【００１１】また、上記話者適応においては少ない発声データから音響モデルを精度良く適応できないために、誤り率を半減させるためには数十単語以上の発声データが必要となり、学習話者に負担を強いることになるという問題がある。さらに、音響モデルの適応に声道長伸縮関数を用いる場合には、上述した話者正規化の場合と同様の問題が発生する。また、音声合成における声質変換の場合にも、同様に少ない発声データからは精度良く声質が得られないという問題がある。

【００１２】そこで、この発明の目的は、少ない発声データから発声データの内容に依存せずに安定して話者特徴を抽出できる話者特徴抽出装置および話者特徴抽出方法、その抽出方法を用いた音声認識装置，音声合成装置、並びに、話者特徴抽出処理プログラムを記録したプログラム記録媒体を提供することにある。

【００１３】

【課題を解決するための手段】上記目的を達成するため、第１の発明は、入力話者の音声から，標準話者の音声スペクトルに対して上記入力話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を話者特徴として抽出する話者特徴抽出装置において、所定の音声単位毎に，上記標準話者の音響モデルに対して，上記入力話者の音声サンプルの尤度あるいは音響モデルを上記入力話者の音声サンプルに話者適応させた話者適応音響モデルの尤度を最大にするという基準に従って，上記周波数伸縮関数を最尤推定し，この推定された上記周波数伸縮関数の集合の頻度分布を求める頻度計測手段と、上記頻度分布に基づいて，最大頻度を有する周波数伸縮関数を話者特徴として抽出するモード抽出手段を備えたことを特徴としている。

【００１４】上記構成によれば、頻度計測手段によって、所定の音声単位毎に、上記周波数伸縮関数の最尤推定が行われ、この推定された周波数伸縮関数の集合の頻度分布が求められる。そして、モード抽出手段によって、最大頻度を有する周波数伸縮関数が話者特徴として抽出される。したがって、上記周波数伸縮関数の頻度分布に複数のピークが存在しても、正確な周波数伸縮関数が安定して抽出される。

【００１５】また、上記第１の発明の話者特徴抽出装置は、入力話者の音声サンプルから音響モデルを用いたビタビアルゴリズムによって音素境界情報を推定する音素境界情報推定手段を備えて、上記頻度計測手段を、上記音素境界情報に基づいて、有音無音の別および調音点の位置に従って、上記最尤推定を行う際に用いる上記入力話者の音声サンプルの音声区間を限定するように成すことが望ましい。

【００１６】上記構成によれば、音素境界情報推定手段で推定された音素境界情報に基づいて、上記頻度計測手段によって、有音無音の別および調音点の位置に従って上記入力話者の音声サンプルの音声区間が限定される。したがって、声道長の差の影響を受け難い音素や無音部を上記最尤推定時における周波数軸伸縮の対象外にして、声道長の差の影響を受け難い音素や無音部まで変形されることが防止可能になる。こうして、少ない発声データから、発声データの内容に依存せずに安定して話者特徴が抽出される。

【００１７】また、上記第１の発明の話者特徴抽出装置は、音響モデルの各状態が，有音無音の別および調音点の位置に従って予め設定された上記最尤推定を行う音声区間に属しているか否かを判別し，属している状態に関して，上記音響モデルを入力音声サンプルに話者適応させて話者適応音響モデルを作成する話者適応モデル作成手段を備えて、上記頻度計測手段を、上記標準話者音響モデルに対して上記話者適応音響モデルを用いて上記最尤推定を行うように成すことが望ましい。

【００１８】上記構成によれば、話者適応モデル作成手段によって、有音無音の別および調音点の位置に従って設定された音声区間に属する音響モデルの状態に関して、入力音声サンプルに話者適応された話者適応音響モデルが作成される。そして、上記頻度計測手段によって、上記標準話者音響モデルに対して上記話者適応音響モデルを用いて上記最尤推定が行われる。したがって、声道長の差の影響を受け難い音素や無音部を上記話者適応音響モデル作成の対象外にして、声道長の差の影響を受け難い音素や無音部までが変形されるのを防止することが可能になる。こうして、少ない発声データから、発声データの内容に依存せずに安定して話者特徴が抽出される。

【００１９】また、上記第１の発明の話者特徴抽出装置は、上記モード抽出手段を、上記最大頻度を有する周波数伸縮関数が複数存在する場合には、上記頻度分布を混合ガウス分布で表現した場合における当該複数の周波数伸縮関数が属している分布の分散が大きい方の周波数伸縮関数をもって話者特徴するように成すことが望ましい。

【００２０】上記構成によれば、上記モード抽出手段によって、上記最大頻度を有する周波数伸縮関数が複数存在する場合には、属している分布の分散が大きい方の周波数伸縮関数が抽出される。こうして、より多くの周波数伸縮関数の集団を代表する周波数伸縮関数が話者特徴として抽出される。

【００２１】また、上記第１の発明の話者特徴抽出装置は、上記モード抽出手段を、上記標準話者の特徴を表す周波数伸縮関数に近い方の周波数伸縮関数をもって話者特徴とする機能を組み合せて、上記話者特徴を抽出るよ

うに成すことが望ましい。

【００２２】上記構成によれば、上記最大頻度を有する複数の周波数伸縮関数が属している分布の分散の大きさが同程度である場合には、より標準話者の周波数伸縮関数に近い方の周波数伸縮関数が、話者特徴として抽出される。

【００２３】また、第２の発明は、標準話者の音声スペクトルに対して入力話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を用いて入力話者の音声スペクトルの周波数軸を伸縮することによって上記入力話者の音声を正規化する正規化手段を有する音声認識装置において、上記正規化手段は、上記第１の発明の話者特徴抽出装置と、上記話者特徴抽出装置によって抽出された周波数伸縮関数を用いて，上記入力話者の音声スペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００２４】上記構成によれば、周波数ワープ手段によって、上記第１の発明の話者特徴抽出装置で抽出された話者の特徴をより的確に表す正確な周波数伸縮関数を用いて話者正規化が行われる。したがって、発声データの内容に依存せずに安定して話者正規化が行われて、高い精度で認識結果が得られる。

【００２５】また、第３の発明は、入力話者の音声スペクトルに対して標準話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を用いて音声のスペクトルの周波数軸を伸縮することによって音響モデルを入力話者に話者適応させる話者適応手段を有する音声認識装置において、上記話者適応手段は、上記第１の発明の話者特徴抽出装置と、上記話者特徴抽出装置によって抽出された周波数伸縮関数の逆関数を用いて，上記音響モデルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００２６】上記構成によれば、周波数ワープ手段によって、上記第１の発明の話者特徴抽出装置で抽出された話者の特徴をより的確に表す正確な周波数伸縮関数の逆関数を用いて、話者適応が行われる。したがって、発声データの内容に依存せずに安定して話者適応が行われて、高い精度で認識結果が得られる。

【００２７】また、第４の発明は、入力話者の音声スペクトルに対して標準話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を用いて音声のスペクトルの周波数軸を伸縮することによって，標準話者の音声素片を接続して成る合成音声の声質を発話者の声質に変換する声質変換手段を有する音声合成装置において、上記声質変換手段は、上記第１の発明の話者特徴抽出装置と、上記話者特徴抽出装置によって抽出された周波数伸縮関数の逆関数を用いて，上記音声素片の周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００２８】上記構成によれば、周波数ワープ手段によって、上記第１の発明の話者特徴抽出装置で抽出された話者の特徴をより的確に表す正確な周波数伸縮関数の逆関数を用いて、声質変換が行われる。したがって、発声データの内容に依存せずに安定して声質変換が行われて、より入力話者の声質に近い合成音声が得られる。

【００２９】また、第５の発明は、入力話者の音声から，標準話者の音声スペクトルに対して上記入力話者の音声スペクトルの周波数軸を伸縮する際の周波数伸縮関数を話者特徴として抽出する話者特徴抽出方法において、所定の音声単位毎に，上記標準話者の音響モデルに対して，上記入力話者の音声サンプルの尤度あるいは音響モデルを上記入力話者の音声サンプルに話者適応させた話者適応音響モデルの尤度を最大にするという基準に従って，上記周波数伸縮関数を最尤推定し、この推定された上記周波数伸縮関数の集合の頻度分布を求め、上記頻度分布に基づいて，最大頻度を有する周波数伸縮関数を話者特徴として抽出することを特徴としている。

【００３０】上記構成によれば、所定の音声単位毎に最尤推定された周波数伸縮関数の集合の頻度分布に基づいて、最大頻度を有する周波数伸縮関数が話者特徴として抽出される。したがって、上記周波数伸縮関数の頻度分布に複数のピークが存在しても、正確な周波数伸縮関数が安定して抽出される。

【００３１】また、第６の発明のプログラム記録媒体は、コンピュータを、上記第１の発明における頻度計測手段およびモード抽出手段として機能させる話者特徴抽出処理プログラムが記録されたことを特徴としている。

【００３２】上記構成によれば、上記第１の発明の場合と同様に、上記周波数伸縮関数の頻度分布に複数のピークが存在する場合でも、正確な周波数伸縮関数が安定して抽出される。

【００３３】

【発明の実施の形態】以下、この発明を図示の実施の形態により詳細に説明する。

＜第１実施の形態＞図１は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者正規化方式を用いた音声認識装置であり、上記ＨＭＭに代表される音響モデルをベースとしている。

【００３４】音声入力部１において、マイクから入力された音声はディジタル波形に変換されて音響分析部２に入力される。音響分析部２は、入力されたディジタル波形を短い時間間隔（フレーム）毎に周波数分析し、スペクトルを表す音響パラメータのベクトル系列に変換する。ここで、上記周波数分析としては、ＭＦＣＣ（メル周波数ＦＦＴ（高速フーリエ変換）ケプストラム）やＬＰＣ（線形予測分析）メルケプストラム等のスペクトルを効率よく表現できる音響パラメータを抽出できる分析方法が用いられる。こうして得られた音響パラメータ系列は、話者正規化部３を構成する周波数ワープ部４に送出される。

【００３５】上記話者正規化部３は、上記周波数ワープ部４と周波数ワープ関数推定部５とから概略構成される。そして、周波数ワープ関数推定部５は、学習時には、音響分析部２からの音響パラメータ系列と単語列入力部６から入力された学習用単語の音素列とに基づいて、音素境界情報および周波数ワーピング関数を推定して周波数ワープ部４に送出する。また、認識時には、音響分析部２からの音響パラメータ系列と上記学習時に推定された周波数ワーピング関数とに基づいて音素境界情報を推定し、この推定された音素境界情報を上記周波数ワーピング関数と共に周波数ワープ部４に送出する。尚、周波数ワープ関数推定部５の構成と動作については後に詳述する。

【００３６】そうすると、上記周波数ワープ部４は、上記周波数ワーピング関数および音素境界情報を用いて、入力音声の音響パラメータ系列を周波数ワープ(話者正規化)し、周波数ワープ後の音響パラメータ系列を尤度演算部７に送出するのである。そして、尤度演算部７では、周波数ワープされた音響パラメータ系列に対して、不特定話者音響モデル格納部８に格納された不特定話者モデル(HMM)を作用させて、各音韻の状態毎に尤度を算出する。そして、得られた尤度系列を照合部９に送出する。

【００３７】上記照合部９は、上記尤度演算部７からの尤度系列に対して、辞書格納部１０に登録された総ての言語モデル(単語)との照合を行ない、各単語のスコアを算出する。そして、上位のスコアを呈する単語を認識候補(認識結果)として出力部１１から出力するのである。

【００３８】以下、上記周波数ワープ関数推定部５の構成と動作について詳細に詳述する。図２に、上記周波数ワープ関数推定部５における学習時に機能する部分の構成を示す。さらに、図３には、周波数ワープ関数推定部５における認識時に機能する部分の構成を示す。先ず、図２に従って、学習時について説明する。

【００３９】音素境界推定部１５は、全話者音響モデル格納部１２に格納された混合数１以上の全話者音響モデル(HMM)を用いて、ビタビアルゴリズムによって音素境界情報を求める。その際に、教師あり学習時には、音素境界推定部１５には、単語列入力部６からの音素列と音響分析部２からの音響パラメータ系列(学習データ)とが入力される。そうすると、音素境界推定部１５は、入力音響パラメータ系列に入力音素列を適用させて、上記全話者音響モデルを用いたビタビアルゴリズムによって音素境界情報を求める。これに対して、教師なし学習時には、音素境界推定部１５には、音響分析部２からの音響パラメータ系列(学習データ)のみが入力される。そうすると、音素境界推定部１５は、入力音響パラメータ系列に言語モデル格納部１３に格納された弱い文法の言語モデルを適用させて、全話者音響モデルを用いたビタビアルゴリズムによって音素境界情報を求めるのである。

そして、こうして得られた音素境界情報は頻度計測部１６および周波数ワープ部４に送出される。

【００４０】尚、上記「弱い文法」とは、対象言語の音素または音節の接続に関する制約条件のみを表現するネットワーク(有限状態オートマトン)のことである。例えば、日本語の場合には、/k/と/i/とは接続するが、/s/と/k/は接続しないというような制約条件である。また、上記音素境界情報とはこの音素境界情報によって分離される音素のラベル情報をも含む概念であり、上記ビタビアルゴリズムによって求まる。

【００４１】上記頻度計測部１６は、入力された音素境界情報に従って、全話者音響モデル格納部１７に格納された混合数１の全話者モデル(HMM)を用いて、後に詳述する方法によって、音響モデルの状態や入力サンプル等の音声単位毎に周波数ワーピング関数ｆの係数αを最尤推定する。さらに、α軸に関する頻度の分布を表す分布関数Ｈ(α)を求める。そして、得られた分布関数Ｈ(α)をモード抽出部１８に送出する。

【００４２】上記モード抽出部１８は、後述するようにして、上記分布関数Ｈ(α)の中から最大頻度を与える最適係数α(つまり周波数ワーピング関数ｆ)を推定する。そして、推定された周波数ワーピング関数ｆを関数格納部１９に格納すると共に、周波数ワープ部４に送出するのである。

【００４３】次に、図３に従って、認識時について説明する。尚、図３における全話者音響モデル格納部１２，言語モデル格納部１３，音素境界推定部１５および関数格納部１９は、図２において学習時に使用される全話者音響モデル格納部１２，言語モデル格納部１３，音素境界推定部１５および関数格納部１９と同じものである。

【００４４】事前ワープ部２０は、上記学習時に推定されて関数格納部１９に格納された周波数ワーピング関数ｆを用いて、認識対象の音響パラメータ系列を周波数ワープする。以下、この場合の周波数ワープを、後に周波数ワープ部４によって行われる周波数ワープに対して「事前ワープ」と言うことにする。こうして、事前ワープが行われた音響パラメータ系列が音素境界推定部１５に送出される。

【００４５】そうすると、上記音素境界推定部１５は、事前ワープが行われた音響パラメータ系列に弱い文法の言語モデルを適用させて、全話者音響モデルを用いたビタビアルゴリズムによって音素境界情報を求めるのである。その場合、認識対象の音響パラメータ系列は、学習時に抽出された話者特徴としての周波数ワーピング関数ｆを用いて事前ワープされている。したがって、より話者の声道長に即した音素境界情報を求めることができるのである。そして、得られた音素境界情報が、関数格納部１９に格納されている周波数ワーピング関数ｆと共に周波数ワープ部４に送出される。

【００４６】そうすると、上記周波数ワープ部４におい

ては、上記推定された周波数ワーピング関数ｆによって、上記学習時には、入力された学習用の音響パラメータ系列が周波数ワープされる。一方、上記認識時には、入力された認識用の学習音響パラメータ系列が周波数ワープされるのである。

【００４７】すなわち、本実施の形態においては、上記学習時における周波数ワープ関数推定部５による係数αの最尤推定を、全音声単位に関して行うのではなく個々の音声単位毎に行い、その最大頻度を呈する係数αを推定することによって、正確な周波数ワーピング関数ｆを安定して推定するのである。また、上記学習時に頻度計測部１６と周波数ワープ部４とによって入力音響パラメータ系列に周波数ワーピング関数ｆを適用する場合、および、認識時に周波数ワープ部４によって入力音響パラメータ系列に周波数ワーピング関数ｆを適用する場合には、後に詳述するように、表１の分類表に従って、上記音素境界情報に基づいて、周波数ワープ(正規化)の対象とする音素区間を限定するのである。こうすることによって高精度認識を行う音声認識装置を構築することができるのである。

【００４８】ところで、上記周波数ワープ関数推定部５における上記周波数ワーピング関数ｆの推定方法には、以下に述べる二通りの推定方法がある。

（Ａ）標準話者の音響モデルを入力音声データに話者適応させた適応モデルを用いる。

（Ｂ）入力音声データを直接用いる。

そして、この二通りの推定方法を、入力音声データの量や質に応じて使い分けるのである。ここで、音声データの質とは尤度の上昇具合であり、周波数ワープ関数推定部５は、上記二通りの推定方法による尤度の上昇具合を見計らって、上昇の大きい推定方法を採用するのである。長いエンロール期間を許容できる音声認識装置の場合には、このような推定処理も可能となる。尚、長いエンロール期間を許容できない場合には、予め何れかの推定方法に固定しておけばよい。

【００４９】上記推定方法(Ａ)は、入力音声データが少ない場合に有効である。また、推定方法(Ｂ)は、入力音声データが多い場合に有効であり、入力音声データから直接求めるために、精密な推定が可能となる。但し、入力音声データが少ない場合には、当該推定をエンロールモードで行う際に、入力音声データに無い音素環境における上記係数の推定や平滑化が問題になる。

【００５０】また、上記推定方法(Ａ),(Ｂ)の各々に関して、使用する音響モデルは、全話者モデルの場合と、話者クラスタ別に作成された混合数が１の音響モデルの場合との二通りがある。音声認識装置の記憶容量が少ない場合には前者を採用する。一方、記憶容量が多い場合は音響モデル群を各話者クラスタ別に格納できるので後者を採用する。後者の場合には、入力音声データに基づいて最適な話者クラスタを選択し、この選択話者クラスタに属する音響モデルを使用することになる。すなわち、図１に示す音声認識装置は、全話者モデルを用いた推定方法(Ｂ)によって上記係数の推定を行うのである。

【００５１】ここで、上記話者クラスタ別に作成された音響モデルとは、ある基準で全学習話者をクラスタリングしておき、複数の話者クラスタ毎に学習によって作成された音響モデルのことである。ここでは、上記クラスタリングの基準として、各話者の声道情報を用いる。尚、周波数ワープ関数推定部５が使用する際には、適切な話者クラスタの音響モデルを選択して用いることになる。

【００５２】次に、上記周波数ワープ関数推定部５が学習時に使用する全話者モデルと、尤度演算部７が上記尤度演算時に使用する不特定話者モデルの作成方法について説明する。上記全話者モデルは、総ての学習話者の音声データを用いて学習した音響モデルである。通常、周波数ワーピング関数ｆの最尤推定に使用する場合には混合数を１に設定する。これに対して、不特定話者モデルは、通常学習話者の音声データをそのまま用いて学習した音響モデルである。しかしながら、本実施の形態のように話者正規化を行う音声認識装置においては、尤度演算部７に入力される音響パラメータ系列は、周波数ワープ部４によって既に正規化されている。したがって、不特定話者モデルも、学習話者の音声データを以下に述べる正規化と同様の手順で正規化した正規化学習データを用いた学習によって作成するのである。その場合、不特定話者に対応させるために、通常では、混合数は１以上に設定される。

【００５３】次に、上記周波数ワープ関数推定部５によって行われる周波数ワーピング関数ｆの推定について説明する。先ず、周波数ワープピング関数ｆの定義について説明する。周波数ワープピング関数ｆ(周波数伸縮関数または単に伸縮関数と言う場合もある)の周波数軸は声道の長さを直接反映しているので声道長伸縮関数とも言う。周波数ワーピング関数ｆは、推定の容易さを考慮して、通常はできるだけ少ないパラメータ数で表現される。本実施の形態においては、周波数ワーピング関数ｆを、以下のようなパラメータが１個からなる区分線形関数であると定義する。

【００５４】周波数ワーピング関数f()：

・$x \leqq \min(\omega/\alpha, \omega)$では、$f(x) = \alpha x$

($\omega \fallingdotseq 4$kHz)　($0.88 < \alpha < 1.13$)

・$\min(\omega/\alpha, \omega) < x$では、

$\alpha > 1$のとき　$f(x) \rightarrow (\omega/\alpha, \omega)$と(fs/2,fs/2)とを結ぶ直線

$\alpha \leqq 1$のとき　$f(x) \rightarrow (\omega, \alpha\omega)$と(fs/2,fs/2)とを結ぶ直線

ここで、α：周波数ワーピング関数f()の係数

fs：サンプリング周波数

尚、上記サンプリング周波数fsは、本実施の形態におい

13

ては８kHz以上を仮定している。すなわち、fs＝１２kHzの場合には、(fs/２,fs/２)は(６kHz, ６kHz)となるのである。また、係数αの定義域「０．８８＜α＜１．１３」は飽くまでも一例であり、子供まで含めると「０．７＜α＜１．１３」となる。α＞１である場合における上述のような折れ線で表される周波数ワーピング関数f(x)を図４に示す。すなわち、周波数ワーピング関数f()の推定とは係数αを推定することである。

【００５５】また、複数のパラメータを有する周波数ワーピング関数f(x)の場合でも、以下に述べる１個のパラメータαを有する周波数ワーピング関数f(x)の場合と同様にして、パラメータ空間の総ての座標における尤度を算出して頻度を計測することによって、音響パラメータ系列に適合した係数を推定することができる。

【００５６】上記周波数ワープ関数推定部５における上記周波数ワーピング関数ｆの推定方法が上記推定方法(A)である場合には、標準話者の音響モデルを入力音声データに話者適応させた適応モデルを用いて、以下の手順によって２つの音響モデルの状態間の尤度を求めて推定するのである。

【００５７】尚、その場合における上記適応モデルは、＊

$$\hat{\alpha}_i = \arg\max_{\alpha} r_i(\bar{\mu}_i^f) \qquad (i \in \Omega) \qquad \cdots(1)$$

ここで、　Ω：正規化対象の出力確率密度関数集合のインデックス

ri()：標準モデルの第ｉ番目の出力確率密度関数

f()：αを係数とする周波数ワーピング関数

$\mu^{-}{}_i{}^f$：入力モデルにおける第ｉ番目の出力確率密度関数ｂｉ()の平均値ベクトル$\mu^{-}{}_i$をf()で周波数ワープしたベクトル

尚、上記正規化対象の出力確率密度関数集合Ωは、後述する正規化対象の音素区間に属する音素に関する音響モデルの出力確率密度関数の集合である。

【００６０】上記音響分析部２による音響分析で得られる音響パラメータや上記音響モデルの出力確率密度関数の引数は、通常ＭＦＣＣやＬＰＣケプストラムである。これらの音響パラメータの各次元はケプストラムと呼ばれる物理量であって、周波数ではない。そこで、上記周波数ワープ処理を行なう際には、学習データである音響パラメータからスペクトルへの変換C^{-1}(ケプストラムの場合は逆cos変換)を行なって周波数次元に変換する。そして、周波数ワープ処理終了後は、逆変換Ｃ(ケプストラムの場合はcos変換)を行なって元の音響パラメータ次元に戻すのである。すなわち、$\mu^{-}{}_i{}^f = C(f(C^{-1}(\mu_i)))$となる。ここで、$C^{-1}$, Ｃは、音響パラメータからスペクトルへの変換とその逆変換である。

【００６１】次に、上記式(１)によって求められた{$\hat{\alpha}_i$}(ｉ∈Ω)に関して、α軸に関する頻度の分布を求め、この頻度分布を表す関数をｈ(α)とおく。そして、上述のごとく正規化対象の出力確率密度関数集合Ωの状態ｉ

14

＊例えば、音響モデルの各状態が予め設定された正規化対象の音素区間に属しているか否かを判別し、属している状態に関して、上記音響モデルを入力音声データに話者適応させる話者適応モデル作成手段によって作成すればよい。

【００５８】ここで、標準モデル(全話者音響モデル格納部１７に格納された全話者モデルに相当)と入力モデル(上記適応モデルに相当)との２つの音響モデルの対応する状態間の尤度を、標準モデルの出力確率密度関数ｒｉ()に、入力モデルの出力確率密度関数ｂｉ()の平均値ベクトル$\mu^{-}{}_i$を上記周波数ワーピング関数f()で周波数ワープして得られたベクトル$\mu^{-}{}_i{}^f$を代入したときの値と定義する。上記各出力確率密度関数は多次元ガウス分布であって、平均値ベクトルと分散ベクトルから成っている。

【００５９】そして、上記正規化対象の出力確率密度関数集合Ωにおける第ｉ番目の状態間の尤度ri($\mu^{-}{}_i{}^f$)に基づいて、状態ｉにおける周波数ワーピング関数f()の最適係数$\hat{\alpha}_i$は、式(１)に示すように尤度ri($\mu^{-}{}_i{}^f$)を最大にする係数として推定されるのである。

毎に最尤推定して得られた係数$\hat{\alpha}_i$のうちモード(並数,最頻値)を与える係数$\hat{\alpha}$を、上記周波数ワーピング関数f()の最適係数として式(２)によって推定するのである。

$$\hat{\alpha} = \arg\max_{\alpha} h(\alpha) \qquad \cdots(2)$$

【００６２】図５に、上記係数αの頻度分布を表す分布関数ｈ(α)の一例を示す。このような多峰性を有する場合、すなわち複数のピークが存在する場合には、従来法によれば２つのピークの中間に最適係数$\hat{\alpha}$が求まる。これに対し、本実施の形態によれば、頻度が高い方のピークを呈するに係数αが最適係数$\hat{\alpha}$として求まるのである。尚、図中、棒グラフは係数αを０．０１きざみで観測した場合の頻度分布ｈ(α)であり、破線はその包絡線である。サンプル数が少ない場合はこの包絡線を分布関数ｈ(α)として差し支えない。

【００６３】ここで、図６に例示するように、上記分布関数ｈ(α)の最大値を与えるαが複数個存在する場合がある。このような場合における最適係数$\hat{\alpha}$の推定は、「係数αが属する分布の分散が大きいこと」および「標準話者に近いこと」の２つの基準を組み合わせて行う。図６においては、モードα_1に属する分布は、分散が小さく、モードα_2に比べて標準値の１．０から離れており、係数αの推定誤りによるゴミと考えられる。

【００６４】今、２つの値α_1, α_2で最大値ｈ(α)が与えられたとする。つまり、ｈ(α_1)＝ｈ(α_2)＝ｈ($\hat{\alpha}$)となる場合である。係数αが属する分布(すなわち、各

15

αkを平均値とするガウス分布)の分散は、分布関数h(α)を混合ガウス分布(αk, σk²)で表現することによって与えられる。混合ガウス分布の推定には、HMMの学習方法であるBaum-Welchアルゴリズムが用いられる。ここで、混合数は最大値を与える個数であり、図６の例の場合は「２」である。尚、αkは平均値であり、σk²は分散であり、図６の場合はｋ＝１，２である。

【００６５】こうした場合、α1を平均値とする分布のα2を平均値とする分布に対する分散の小ささの度合いは、例えば式(３)で表される。

$$s(1,2)=\sigma_2^2-\sigma_1^2 \quad \cdots(3)$$

また、標準話者の係数αはα＝１であるから、αkの標準話者への近さは例えば式(４)で表される。

$$d_k=|\alpha_k-1| \quad \cdots(4)$$

そして、通常は、式(３)におけるｓ(１，２)の値に基づいて最適係数α^を選択する。そして、ｓ(１，２)の値が非常に小さく、両者の分散が同程度と見なされる場合には、式(４)におけるｄkの値に基づいて最適係数α^を選択するのである。

【００６６】または、以下のようにして最適係数α^を選択してもよい。すなわち、αk1のαk2に対するスコアｇ(k1, k2)を式(５)で定義する。そして、このスコアの値をα1とα2とについて求め、小さい方を最適係数α^として選択するのである。

$$g(k_1,k_2)=\lambda * s(k_1,k_2)+(1-\lambda) * d_{k1} \quad \cdots(5)$$

ここで、λはシミュレーション実験に基づいて与えられる重み係数であり、[０，１]の間、例えば０．７等に設定される。

【００６７】このように、本実施の形態においては、従来のように、全状態に関して係数αを最尤推定するのではなく、個々の状態ｉ毎に最尤推定して最大頻度を呈する係数αを求めるのである。こうすることによって、各状態ｉ毎の係数αiの集合における頻度分布に複数のピークが存在する場合でも、正確な周波数ワーピング関数ｆ()を安定して推定できるのである。また、その際における各状態ｉ毎の係数αの最尤推定を、上記正規化対象の音素区間に属する音素に関してのみ行うことによって、少ない音声データによって、精度良く周波数ワーピング関数f()を推定できるのである。

【００６８】一方、上記周波数ワープ関数推定部５における上記周波数ワーピング関数ｆの推定方法が上記推定方法(B)である場合には、音響分析部２からの入力音響パラメータ系列を直接用いて、以下の手順によって周波数ワーピング関数f()の最適係数αを推定する。尚、上述したごとく、図１に示す音声認識装置における周波数ワープ関数推定部５には、上記推定方法(B)が適用されている。したがって、以下の推定手順を行うことになる。

$$\bar{\alpha}_j=\arg\max_{\alpha} P(X_j^f|W_j) \quad (j\in\Psi) \quad \cdots(6)$$

上記(１)～(４)の処理を総ての入力音声サンプル{Xj}(ｊ∈Ψ)に対して実行して、各サンプルXj毎の最適係数にα¯jを求める。以上の手順(１)から手順(４)までの処理は、頻度計測部１６によって行われる。

【００７３】(５) 上記求められたα¯jを係数とする周波数ワーピング関数f¯j()を用いて、ビタビアルゴリズムによって、各サンプル毎に音素境界情報が求められる。そして、全入力音声サンプル{Xj}(ｊ∈Ψ)のうち、各サンプル毎の音素境界情報に基づく正規化対象と★なる音素区間の音響パラメータ系列の集合を{X¯j}とおく。この手順(５)による音素境界情報算出処理は音素境界推定部１５で行われる。

$$\hat{\alpha}_j=\arg\max_{\alpha} P(\bar{X}_j^f|W_j) \quad (j\in\Psi) \quad \cdots(7)$$

【００７４】(６) 上記正規化対象となる音素区間の音響パラメータ系列の集合{X¯j}に関して、上記音素境界情報に基づく正規化対象の音素区間毎に、ビタビアルゴリズムによって累積尤度Ｐ(X¯jf|Wj)を求める。そして、式(７)によって、最適係数α^jが求め直される。

【００７５】(７) 全サンプルの正規化対象音素区間における{α^j}に関して頻度分布を求め、頻度分布を表す分布関数をＨ(α)とおく。そして、上述した推定方法(A)の場合と同様に、係数α^jのうちモード(並数，最頻

値)を与える係数α^を、上記周波数ワーピング関数f()の最適係数として式(8)によって推定するのである。

$$\hat{\alpha} = \arg\max_{\alpha} H(\alpha) \quad \cdots(8)$$

上記手順(6)とこの手順(7)における頻度分布の算出とは頻度計測部１６によって行われる。また、手順(7)における上記モードを与える係数α^の抽出はモード抽出部１８によって行われる。

【００７６】尚、上記頻度分布に同一値のピークが複数の存在する場合には、上記推定方法(A)において述べた方法と同様の方法によって最適係数を推定する。

【００７７】このように、上記推定方法(B)の場合には、個々のサンプルｊ毎に係数αｊを最尤推定し、最大頻度を呈するαを最適係数として求めることによって、各サンプルｊ毎の係数αｊの集合における頻度分布に複数のピークが存在する場合でも正確な周波数ワーピング関数ｆ()を安定して推定できるのである。また、その際における各サンプルｊ毎の係数αｊの最尤推定を、上記正規化対象の音素区間に属する音素に関してのみ行うことによって、少ない音声データによって、精度良く周波数ワーピング関数f()を推定できるのである。

【００７８】尚、上記周波数ワーピング関数f()推定処理における音素境界推定部１５と頻度計測部１６とモード抽出部１８との処理の区分は、上述に限定されるものではない。例えば、手順(2)および手順(6)におけるビ＊

＊タビ演算を、音素境界推定部１５で行うようにしても差し支えない。

【００７９】次に、上記学習時には頻度計測部１６と周波数ワープ部４とで、認識時には周波数ワープ部４で周波数ワープを行う際に、頻度計測部１６および周波数ワープ部４によって行われる上記音素境界情報に基づく対象音素区間の限定について説明する。

【００８０】上述したように、学習時および認識時においては、周波数ワープ関数推定部５の音素境界推定部１５によって、入力話者の音響パラメータ系列あるいはこの入力音響パラメータ系列に基づく適応モデルに、発話内容の音素列や言語モデル格納部１３に格納された弱い文法の言語モデルを適用させて、全話者音響モデル格納部１２に格納された全話者モデルや話者クラスタにクラスタリングされた全話者モデルから選択されたものを用いたビタビアルゴリズムによって音素境界情報を求め、頻度計測部１６(学習時)および周波数ワープ部４(学習時,認識時)に送出するようにしている。

【００８１】そうすると、上記頻度計測部１６および周波数ワープ部４は、上記周波数ワープ関数推定部５からの音素境界情報に基づいて、入力音声データのうち周波数ワープ処理の対象とする音素区間を制御するのである。本実施の形態においては、音素を表１に示す５種類に分類する。表１

分類	音素	説明
[a]	無音区間	環境騒音そのものであり、声道長とは全く関係ない。
[b]	調音点が歯茎より前に位置する子音	音源が声道出口付近に位置するために、声道全体による共鳴管が形成されず、声道長の影響を受け難い。
[c]	調音点が歯茎より後に位置する子音,半母音	母音と同様に声道長に直接影響される。しかしながら、子音や半母音区間は元来安定していないので声道長の推定には適さない。
[d]	「ウ」を除く母音	母音は全般的に声道長の影響を直接受ける。
[e]	母音「ウ」,撥音	日本語の「ウ」は発声の仕方によってフォルマンと周波数が大きく変動するので、声道長の影響よりも発声の仕方に大きく影響される。撥音も音素環境に大きく依存すると共に鼻音化の影響が大きい。

【００８２】そして、この分類に基づいて、以下のような区別に従って、上記頻度計測部１６は学習時の周波数ワープを制御し、周波数ワープ部４は学習時および認識時の正規化を制御するのである。

・学習時…分類[ｄ]

・認識時…分類[ｃ],分類[ｄ],分類[ｅ],(分類[ｂ])

但し、認識時には、分類[ｂ]を含めてもよい。発音の仕方によっては、音素「イ」も音素「ウ」と同様に狭母音なのでフォルマント周波数が大きく変動する場合がある。したがって分類[ｅ]に音素「イ」を含め、分類[ｄ]から音素「イ」を除いてもよい。

【００８３】尚、上記周波数ワープ部４による正規化処理対象の音素区間制御方法は、周波数ワープ関数推定部５の音素境界推定部１５が用いる全話者モデルの規模に応じて二通りある。

・全話者音響モデル格納部１２の容量に余裕があるために、全話者モデルの規模を非常に大きくできる場合には、分類[ｂ]の調音点が歯茎より前に位置する子音を分離可能な音素境界情報を精度良く推定できるので、分類[ｃ],分類[ｄ],分類[ｅ]のみを正規化対象区間とする。

・全話者モデルの規模をある程度大きくできる場合には、分類[ｂ]を分離可能な音素境界情報を推定できないために上述のごとく分類[ｂ]を入れて、分類[ｂ],分類[ｃ],分類[ｄ],分類[ｅ]を正規化対象区間とする。つまり、無音区間のみを正規化対象の音素から外すのである。

19

【００８４】上述したように、本実施の形態における音声認識装置は、高精度認識を行うために周波数ワープ部４において周波数ワープの対象とする音素区間を限定するようにしている。しかしながら、計算資源(処理能力)に余裕がないシステムに搭載する場合には、全話者モデルの規模を大きくできないため精度良く音素境界情報を推定することができない。そのような場合には、周波数ワープ部４を常に動作させて、全音素区間を対象に周波数ワープを行っても差し支えない。このように精度良く音素境界情報を推定できない場合でも、分類[ｄ]の声道長の影響を直接受ける母音は推定できる。したがって、周波数ワープ関数推定部５の頻度計測部１６によって推＊

20

＊定された周波数ワーピング関数ｆ()は、音素境界推定部１５からの音素境界情報に基づいて声道長の影響を直接受ける分類[ｄ]の母音のみから得られていることになる。したがって、周波数ワープ部４による周波数ワープの際に声道長の影響を受け難い音素区間と無音区間とが不必要に変形されることを防止するという効果は得ることができるのである。

【００８５】最後に、上記周波数ワープ関数推定部５の音素境界推定部１５が、学習時および認識時に用いる言語モデルについて説明する。表２に、各動作モード時における周波数ワープ関数推定部５が用いる言語モデルの切換状況を示す。表２

動作モード		言語モデル
学習	教師あり	単語列入力部６からの音素列
	教師なし	弱い文法の言語モデルまたは認識結果
認識	高精度	弱い文法の言語モデル
	通常	なし(不使用)

【００８６】表２において、通常の認識処理時における言語モデル「なし」とは、上述のごとく全音素区間を正規化対象とするために正規化対象制御用の音素境界情報を推定する必要がなく、ビタビアルゴリズムを動作させないために言語モデルを使用しないという意味である。また、学習モードにおける「教師あり」とは、上述したように、音素境界情報の推定時にビタビアルゴリズムを行う際に発話内容の音素列を使用することであり、単語列入力部６から入力される音素列そのものが言語モデルとなる。これに対して、「教師なし」とは、発話内容の音素列を使用しないものであり、言語モデル格納部１３に格納された弱い文法の言語モデルを使用するのである。

【００８７】尚、上記弱い文法の言語モデルに代えて、認識結果を使用することも可能である。この場合、照合部９からの出力である認識結果を発話内容の音素列(言語モデル)として使用するのである。つまり、一度認識処理を行ってから再び学習モード時における周波数ワープ関数推定部５の処理動作に戻るのである。その場合には、図１に破線で示すように、出力部１１からの認識単語列を一種の教師音素列として単語列入力部６に入力する。但し、発話内容に規制が無いので照合部９用の言語モデルを、音素境界推定部１５でのビタビ演算に流用してよいかどうかという問題はある。

【００８８】以上、上記HMMに代表される音響モデルを用いた音声認識装置について述べてきたが、音声波形または音声パラメータ系列を標準パターンとして登録しておくタイプの音声認識装置においても、入力音声データを直接用いる推定方法(Ｂ)の場合と同様の手法によってサンプル毎の頻度を観測することによって、本実施の形態における話者正規化方法を適用することができる。

$$\tilde{\alpha}_j = \arg\min_{\alpha} d(X_j^f \mid W_j)$$

尚、その場合には、尤度の代わりにスペクトル間の距離尺度を用いることになる。処理手順は以下の通りである。

【００８９】(１) αに初期値を代入する。入力音響パラメータ系列を$X_j=\{x^-_j(t)\}$ ($t=1,2,\cdots,T_j$)とおく。また、それに対応する標準パターンの音響パラメータ系列を$R_j=\{r^-_j(t)\}$ ($t=1,2,\cdots,T'_j$)とおく。なお、「$x^-_j(t)$」,「$r^-_j(t)$」は時刻(フレーム)ｔにおける音響パラメータベクトル、ｊは各パターン(上記標準パターンに対応)のインデックス、「T_j」,「T'_j」は音響パラメータ系列X_j, R_jにおける最終時刻(最終フレーム)である。

【００９０】(２) 上記入力音響パラメータ系列X_jに、αを係数とする周波数ワーピング関数ｆ()を作用させて周波数ワープを行う。そして、周波数ワープの結果をX_j^fとおく。

【００９１】(３) 上記周波数ワープ後の入力音響パラメータ系列X_j^fと標準パターンの音響パラメータ系列R_jとの累積距離ｄ(X_j^f, R_j)を、ＤＰマッチングによって求める。尚、上記ＤＰマッチングにおいては、距離尺度としてケプストラム距離等のスペクトル間距離を用いる。

【００９２】(４) 係数αを、定義域「０．８８＜α＜１．１３」内において、例えば０．０２きざみで移動させながら、上記(２)と(３)との処理とを繰り返して累積尤度ｄを求める。

【００９３】(５) 上記累積尤度ｄを最小とするα$_j$を求め、これをα^-_jとおく。すなわち、α^-_jは、式(９)によって表される。

$$\cdots(9)$$

21

上記(1)～(5)の処理を総ての入力音声サンプル{Xj}(j∈Ψ)に対して実行して、各サンプルXj毎の最適係数に$\bar{\alpha}_j$を求める。

【００９４】(6) 上記求められた総ての$\bar{\alpha}_j$に関して頻度分布を求め、この頻度分布を表す分布関数をH(α)＊

$$\hat{\alpha} = \arg\max_{\alpha} H(\alpha)$$

【００９５】ここで、上記頻度分布に同一値のピークが複数の存在する場合には、上記推定方法(A)において述べた方法と同様の方法によって最適係数を推定する。

【００９６】尚、サブワードHMMとは異なり、本例における上記標準パターンに音素情報は含まれていない。その場合には、音素等による正規化対象区間の制御は困難であるため導入はしない。その代わり、学習時には、一つのサンプルが単母音や特に表１における分類[ｄ]に相当する母音で成る学習データを入力させるようにすることによって、正規化対象区間の制御を行えばよい。

【００９７】上述したように、本実施の形態においては、上記音素境界推定部１５,頻度計測部１６およびモード抽出部１８を有する周波数関数推定部５を備えている。そして、音素境界推定部１５は、学習時には、音響分析部２からの入力音響パラメータ系列に、教師ありの場合には単語列入力部６からの音素列(言語モデル)を適用させる一方、教師なしの場合には言語モデル格納部１３に格納された弱い文法の言語モデルを適用させて、全話者音響モデル格納部１２に格納された全話者音響モデルを用いたビタビアルゴリズムによって音素境界情報を求める。

【００９８】そうすると、上記頻度計測部１６は、各サンプルｊ毎に、上記周波数ワーピング関数ｆ()の係数αを定義域において所定値ずつ増加させながら入力音響パラメータ系列Xjの周波数ワープを行う。そして、周波数ワープが行われた入力音響パラメータ系列Xj^{f}のうち、上記音素境界情報に基づいて上記表１に従って上述のように設定された正規化対象区間の音響パラメータ系列のみに関して、全話者音響モデル格納部１７に格納された全話者モデルに対する累積尤度Ｐを最大にする係数$\hat{\alpha}_j$を最尤推定する。そして、各サンプルｊ毎の{$\hat{\alpha}_j$}に関する頻度分布を表す分布関数Ｈ(α)を求める。

【００９９】さらに、上記モード抽出部１８によって、係数$\hat{\alpha}_j$のうち最頻値を与える係数$\hat{\alpha}$が周波数ワーピング関数ｆ()の最適係数として推定され、この最適係数$\hat{\alpha}$を係数とする周波数ワーピング関数ｆ()を関数格納部１９に格納するのである。

【０１００】これに対して、認識時には、上記音素境界推定部１５によって、上記教師なし学習時と同様に弱い文法の言語モデルを適用させて、ビタビアルゴリズムによって音素境界情報を求めるのである。

【０１０１】こうして、上記周波数関数推定部５によって、推定された周波数ワーピング関数ｆ()と音素境界情

22

＊とおく。そして、上述した推定方法(A),(B)の場合と同様に、係数$\bar{\alpha}_j$のうち上記モードを与える係数$\hat{\alpha}$を上記周波数ワーピング関数ｆ()の最適係数として式(10)によって推定するのである。

$$\cdots(10)$$

報とが周波数ワープ部４に送出される。そして、周波数ワープ部４によって、上記音素境界情報に基づいて正規化対象となる音素区間が上記表１に従って学習時および認識時に応じて上述のように制御され、その制御結果に従って、当該認識対象の入力音響パラメータ系列が周波数ワープされるのである。

【０１０２】したがって、本実施の形態によれば、話者と標準話者との声道長の差を表わす声道長正規化係数αを係数とする周波数ワーピング関数ｆ()を用いて、頻度計測部１６および周波数ワープ部４によって入力音響パラメータ系列を周波数ワープ(正規化)するに際して、周波数ワープの対象となる音素区間を制御することができる。その結果、声道長の差の影響を受け難い音素や無音部を正規化対象外とすることによって、声道長の差の影響を受け難い音素や無音部まで学習および正規化されてしまうことを防止できる。

【０１０３】さらに、上記学習時における周波数ワープ関数推定部５による係数αの最尤推定を、全サンプル(推定方法(A)の場合には状態)に関して行うのではなく個々のサンプル(または状態)毎に行い、その頻度分布における最大頻度を呈する係数αをもって周波数ワーピング関数ｆの最適係数としている。したがって、上記頻度分布に複数のピークが存在する場合でも、正確な周波数ワーピング関数ｆを安定して推定できるのである。

【０１０４】すなわち、本実施の形態においては、少ない発声データから安定して話者特徴を抽出し、その抽出結果を用いて精度よく話者正規化することによって、高い認識性能を得ることができるのである。

【０１０５】また、上記実施の形態においては、上記係数αの分布関数ｈ(α)に最大値を与えるαがα$_1$とα$_2$との２個存在する場合には、α$_1$を平均値とする分布のα$_2$を平均値とする分布に対する分散の小ささの度合いを式(3)で求め、α$_k$の標準話者への近さを式(4)で求める。そして、「係数αが属する分布の分散が大きいこと」および「標準話者に近いこと」の２つの基準を組み合わせて、上記最適係数αの推定を行うようにしている。したがって、同一最大ピーク値が複数存在するような分布関数ｈ(α)が得られた場合でも、安定して上記最適係数αを推定することができるのである。

【０１０６】また、上記周波数ワープ部４による正規化対象となる音素区間の制御は、上記表１の音素分類に従って、学習時には分類[ｄ](「ウ」を除く母音)を正規化対象音素区間とする。さらに、 認識時には分類[ｃ](調音

23

点が歯茎より後に位置する子音，半母音），分類［ｄ］，分類［ｅ］（母音「ウ」，撥音），（分類［ｂ］（調音点が歯茎より前に位置する子音））を正規化対象音素区間とするようにしている。こうして、学習時および認識時における非正規化音素区間を、有音無音の別および調音点の位置に従って設定することによって、声道長の影響を受け難い音素区間と無音区間とが学習および正規化されることを、確実に防止することができるのである。

【０１０７】＜第２実施の形態＞図７は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者適応方式を用いた音声認識装置である。音声入力部２１，音響分析部２２，単語列入力部２６，尤度演算部２７，照合部２９，辞書格納部３０および出力部３１は、図１に示す上記第１実施の形態における音声入力部１，音響分析部２，単語列入力部６，尤度演算部７，照合部９，辞書格納部１０および出力部１１と同様である。また、周波数ワープ関数推定部２４，全話者音響モデル格納部３２，言語モデル格納部３３および不特定話者音響モデル格納部３４は、図１に示す周波数ワープ関数推定部５，全話者音響モデル格納部１２，言語モデル格納部１３および不特定話者音響モデル格納部８と同様である。尚、周波数ワープ関数推定部２４，全話者音響モデル格納部３２，言語モデル格納部３３および不特定話者音響モデル格納部３４は、周波数ワープ部２５と共に、話者適応部２３を構成している。

【０１０８】上記話者適応部２３の周波数ワープ関数推定部２４は、上記第１実施の形態の場合と同様にして、学習音響パラメータ系列に発話内容の音素列または弱い文法の言語モデルを適用して、全話者モデルを用いたビタビアルゴリズムを行って、音素境界情報および周波数ワーピング関数f()を推定する。そうすると、周波数ワープ部２５は、この推定された周波数ワーピング関数f()の逆関数を用いて、不特定話者音響モデル格納部３４に格納された不特定話者モデルを周波数ワープする。その場合、上記周波数ワープに際しては、上記音素境界情報に基づいて、上記表１における分類［ｂ］，分類［ｃ］，分類［ｄ］，分類［ｅ］に該当する音素の状態に対してのみ変換を行うことによって行う。そして、それ以外の状態は変換しないのである。但し、声道長の影響を受け難い分類［ｂ］に該当する音素の状態は、変換しない場合もある。こうして周波数ワープされた不特定話者音響モデルを、話者適応モデル（HMM）として話者適応音響モデル格納部２８に格納するのである。

【０１０９】こうして学習が終了すると、認識時には、上記尤度演算部２７によって、音響分析部２２からの入力音声の音響パラメータ系列に対して、話者適応音響モデル格納部２８に格納された話者適応モデルを作用させて、上述した尤度演算処理を行なうのである。

【０１１０】このように、本実施の形態においては、学習時に、上記周波数ワープ関数推定部２４によって、学

24

習音響パラメータ系列に基づいて上記音素境界情報および周波数ワーピング関数f()を推定する。そして、周波数ワープ部２５によって、上記推定された周波数ワーピング関数f()の逆関数を用いて、分類［ｃ］，分類［ｄ］，分類［ｅ］（，分類［ｂ］）に該当する音素の不特定話者モデルを周波数ワープすることによって、不特定話者モデルを話者適応させるようにしている。

【０１１１】したがって、本実施の形態によれば、上記不特定話者モデルを話者適応させる際における非正規化音素区間を、無音区間と長音点が歯茎より前に位置する子音とに設定することができる。その結果、声道長の影響を受け難い音素区間と無音区間とが不必要に変形されることを確実に防止することができるのである。

【０１１２】さらに、上記学習時における周波数ワーピング関数f()の推定に際して係数αの最尤推定を個々の状態やサンプル毎に行い、その最大頻度を呈する係数αをもって周波数ワーピング関数ｆの最適係数としている。したがって、各状態やサンプル毎の係数αの集合における頻度分布に複数のピークが存在する場合でも、正確な周波数ワーピング関数ｆを安定して推定することができる。

【０１１３】すなわち、本実施の形態によれば、少ない発声データから安定して話者特徴を抽出し、その抽出結果を用いて精度よく話者適応を行うことによって、高い認識性能を得ることができるのである。

【０１１４】尚、本実施の形態における上記話者適応音響モデル格納部２８に格納する話者適応モデルの与え方には、上述の与え方の以外に、話者クラスタを用いる方法を採用してもよい。そして、この二通りの与え方を、音声認識装置の規模や入力音声データの量や質に応じて使い分けるのである。ここで、音声データの質とは尤度の上昇具合であり、話者適応部２３は、上記二通りの与え方による尤度の上昇具合を見計らって、上昇の大きい推定方法を採用するのである。長いエンロール期間が許容できる音声認識装置の場合には、このような推定処理も可能となる。尚、上記話者クラスタを用いる方法においては、学習音声データに対する尤度が最大値になる話者クラスタの音響モデルを選択する。そして、この選択された音響モデルを話者適応モデルとして話者適応音響モデル格納部２８に格納するのである。

【０１１５】また、上述した二つの与え方の何れかによって得られた話者適応モデルを初期モデルとして、上記ＭＬＬＲ方やＶＦＳ法等の既存の話者適応技術を用いて話者適応を行って新たに話者適応モデルを生成し、これを尤度演算部で用いるようにしても差し支えない。

【０１１６】＜第３実施の形態＞図８は、本実施の形態のテキスト音声合成装置におけるブロック図である。なお、このテキスト音声合成装置は、声質変換方式を用いたテキスト音声合成装置である。テキスト解析部４１は、単語とそのアクセント型とが格納されたアクセント

辞書４２を用い、入力テキストに対して形態素解析および係り受け解析を行って音素文字列とアクセント情報とを生成して韻律生成部４３に送出する。韻律生成部４３は、韻律制御テーブル４４を参照して、継続時間長やピッチやパワーの韻律情報を生成して、音素文字列と共に音声素片選択部４５に送出する。そうすると、音声素片選択部４５は、音声素片辞書４６から音素環境や韻律環境に最適な音声素片を選択し、音声素片情報を生成する。そして、この生成された音声素片情報を周波数ワープ部４８に出力する一方、上記韻律情報を音声素片合成部４７に出力する。

【０１１７】一方、周波数ワープ関数推定部４９は、声質変換のターゲット話者の入力音声波形を基に、第１，第２実施の形態の場合と同様にして、上記音素境界情報および周波数ワーピング関数f()を推定する。そうすると、周波数ワープ部４８は、音声素片選択部４５からの音声素片情報に含まれた音素境界情報に基づいて音質変換対象となる音素区間を上記表１に従って上述のように選択する。そして、その選択結果に従って、当該音質変換対象の音声素片情報である音響パラメータ系列を、上記推定された周波数ワーピング関数f()の逆関数を用いて周波数ワープし、周波数ワープ後の音声素片情報を音声素片合成部４７に送出する。最後に、音声素片合成部４７は、周波数ワープ部４８からの周波数ワープ後の音声素片情報(音声素片の音響パラメータ系列)と音声素片選択部４５からの韻律情報とを用いて、音声波形を生成しスピーカ５０から音声出力するのである。

【０１１８】上述のように、本実施の形態においては、テキスト音声合成を行うに際して、上記周波数ワープ関数推定部４９によって、声質変換のターゲット話者における入力音声の音響パラメータ系列から上記音素境界情報および周波数ワーピング関数f()を推定する。そして、周波数ワープ部４８によって、上記音声素片情報に含まれた音素境界情報に基づいて音質変換対象となる音素区間を制御し、上記推定周波数ワーピング関数f()の逆関数を用いて、テキストに基づいて選択された音声素片の音質変換対象となる音響パラメータ系列を周波数ワープすることによって、声質変換を行うようにしている。

【０１１９】したがって、本実施の形態によれば、テキストに基づいて選択された音声素片をターゲット話者の音質に変換する際における非声質変換音素区間を、無音区間と長音点が歯茎より前に位置する子音とに設定することができる。その結果、声道長の影響を受け難い音素区間と無音区間とが不必要に変形されることを確実に防止することができるのである。

【０１２０】さらに、上記学習時における周波数ワーピング関数f()の推定に際して係数αの最尤推定を個々の状態やサンプル毎に行い、その最大頻度を呈する係数αをもって周波数ワーピング関数ｆの最適係数としている。したがって、各状態やサンプル毎の係数αの集合における頻度分布に複数のピークが存在する場合でも、正確な周波数ワーピング関数ｆを安定して推定することができる。

【０１２１】すなわち、本実施の形態によれば、少ない発声データから安定して話者特徴を抽出し、その抽出結果を用いて精度よく声質変換を行うことによって正しく音質変換を行うことができるのである。

【０１２２】本実施の形態はスペクトル包絡の変換であり、声質の適応におおいに効果がある。しかしながら、話者間の声の特徴差は声質だけでなはく韻律が大きく寄与する。したがって、本実施の形態に対して韻律の適応技術を併用しても構わない。

【０１２３】尚、上述した各実施の形態においては、上記周波数ワープ部４，２５，４８において音響パラメータ系列を周波数ワープする場合に、音声素片選択部４５からの音声素片情報に含まれた音素境界情報に基づいて周波数ワープの対象となる音素区間を制御するようにしている。しかしながら、この発明においては、必ずしもその必要はなく、総ての音素区間に対して周波数ワープを行っても構わない。その場合であっても、周波数ワープ関数推定部５，２４，４９によって推定された周波数ワーピング関数f()は、上記音素境界推定部１５からの音素境界情報に基づいて声道長の影響を直接受ける分類[ｄ]の母音のみから推定されている。したがって、周波数ワープ部４，２５，４８による周波数ワープの際に声道長の影響を受け難い音素区間と無音区間とが不必要に変形されることを防止するという効果は得ることができるのである。

【０１２４】また、上述した各実施の形態においては、上記周波数ワーピング関数f()で成る話者特徴を用いて話者正規化または話者適応を行う音声認識装置、および、上記周波数ワーピング関数f()で成る話者特徴を用いて声質変換を行う音声合成装置について説明している。しかしながら、この発明は、上記周波数ワーピング関数f()を話者特徴として抽出する話者特徴抽出装置にも適用されるものである。

【０１２５】ところで、その場合の話者特徴抽出装置における上記頻度計測手段およびモード抽出手段としての機能は、プログラム記録媒体に記録された話者特徴抽出処理プログラムによって実現される。上記プログラム記録媒体は、ＲＯＭ(リード･オンリ･メモリ)でなるプログラムメディアである。あるいは、外部補助記憶装置に装着されて読み出されるプログラムメディアであってもよい。尚、何れの場合においても、上記プログラムメディアから話者特徴抽出処理プログラムを読み出すプログラム読み出し手段は、上記プログラムメディアに直接アクセスして読み出す構成を有していてもよいし、ＲＡＭ(ランダム･アクセス･メモリ)に設けられたプログラム記憶エリア(図示せず)にダウンロードして、上記プログラ

ム記憶エリアにアクセスして読み出す構成を有していてもよい。尚、上記プログラムメディアからＲＡＭの上記プログラム記憶エリアにダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。

【０１２６】ここで、上記プログラムメディアとは、本体側と分離可能に構成され、磁気テープやカセットテープ等のテープ系、フロッピー（登録商標）ディスク，ハードディスク等の磁気ディスクやＣＤ（コンパクトディスク）－ＲＯＭ，ＭＯ（光磁気）ディスク，ＭＤ（ミニディスク），ＤＶＤ（ディジタルビデオディスク）等の光ディスクのディスク系、ＩＣ（集積回路）カードや光カード等のカード系、マスクＲＯＭ，ＥＰＲＯＭ（紫外線消去型ＲＯＭ），ＥＥＰＲＯＭ（電気的消去型ＲＯＭ），フラッシュＲＯＭ等の半導体メモリ系を含めた、固定的にプログラムを坦持する媒体である。

【０１２７】また、上記各実施の形態における音声認識装置，音声合成装置および話者特徴抽出装置は、モデムを備えてインターネットを含む通信ネットワークと接続可能な構成を有していれば、上記プログラムメディアは、通信ネットワークからのダウンロード等によって流動的にプログラムを坦持する媒体であっても差し支えない。尚、その場合における上記通信ネットワークからダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。または、別の記録媒体からインストールされるものとする。

【０１２８】尚、上記記録媒体に記録されるものはプログラムのみに限定されるものではなく、データも記録することが可能である。

【０１２９】

【発明の効果】以上より明らかなように、第１の発明の話者特徴抽出装置は、頻度計測手段によって、所定の音声単位毎に周波数伸縮関数を最尤推定して頻度分布を求め、モード抽出手段によって最大頻度を有する周波数伸縮関数を話者特徴として抽出するので、上記周波数伸縮関数の頻度分布に複数のピークが存在する場合でも、正確な周波数伸縮関数を安定して抽出することができる。したがって、この発明によれば、発声データの内容に依存せずに安定して話者特徴を抽出できる。

【０１３０】また、上記第１の発明の話者特徴抽出装置は、入力話者の音声サンプルから音素境界情報を推定する音素境界情報推定手段を備えて、上記頻度計測手段を、上記音素境界情報に基づいて、有音無音の別および調音点の位置に従って、上記最尤推定を行う際に用いる上記入力話者の音声サンプルの音声区間を限定するように成せば、声道長の差の影響を受け難い音素や無音部を上記最尤推定時における周波数軸伸縮の対象外にして、声道長の差の影響を受け難い音素や無音部まで変形されることを防止することが可能になる。したがって、少ない発声データから発声データの内容に依存せずに安定して話者特徴を抽出できる。

【０１３１】また、上記第１の発明の話者特徴抽出装置は、音響モデルの各状態が、有音無音の別及び調音点の位置に従って予め設定された上記最尤推定を行う音声区間に属しているか否かを判別し、属している状態に関して、上記音響モデルを入力音声サンプルに話者適応させて話者適応音響モデルを作成する話者適応モデル作成手段を備えて、上記頻度計測手段を、標準話者音響モデルに対して上記話者適応音響モデルを用いて上記最尤推定を行うように成せば、声道長の差の影響を受け難い音素や無音部を上記話者適応音響モデル作成の対象外にして、声道長の差の影響を受け難い音素や無音部まで変形されることを防止することが可能になる。したがって、少ない発声データから発声データの内容に依存せずに安定して話者特徴を抽出できる。

【０１３２】また、上記第１の発明の話者特徴抽出装置は、上記モード抽出手段を、上記最大頻度を有する周波数伸縮関数が複数存在する場合には、上記頻度分布を混合ガウス分布で表現した場合における当該複数の周波数伸縮関数が属している分布の分散が大きい方の周波数伸縮関数をもって話者特徴するように成せば、より多くの周波数伸縮関数の集団を代表する周波数伸縮関数を話者特徴として抽出することができる。

【０１３３】また、上記第１の発明の話者特徴抽出装置は、上記モード抽出手段を、上記標準話者の特徴を表す周波数伸縮関数に近い方の周波数伸縮関数をもって話者特徴とする機能を組み合せて、上記話者特徴を抽出るように成せば、上記最大頻度を有する複数の周波数伸縮関数が属している分布の分散の大きさが同程度である場合でも、より適切な周波数伸縮関数を話者特徴として抽出することができる。

【０１３４】また、第２の発明の音声認識装置は、正規化手段を、上記第１の発明の話者特徴抽出装置と、上記話者特徴抽出装置によって抽出された周波数伸縮関数を用いて入力話者の音声スペクトルの周波数軸を伸縮する周波数ワープ手段で構成したので、話者の特徴をより的確に表す正確な周波数伸縮関数を用いて話者正規化を行うことができる。したがって、発声データの内容に依存せずに安定して話者正規化を行って高い精度で認識結果を得ることができる。

【０１３５】また、第３の発明の音声認識装置は、話者適応手段を、上記第１の発明の話者特徴抽出装置と、上記話者特徴抽出装置で抽出された周波数伸縮関数の逆関数を用いて音響モデルの周波数軸を伸縮する周波数ワープ手段で構成したので、話者の特徴をより的確に表す正確な周波数伸縮関数を用いて話者適応を行うことができる。したがって、発声データの内容に依存せずに安定して話者適応を行って高い精度で認識結果を得ることができる。

【０１３６】また、第４の発明の音声合成装置は、声質

29

変換手段を、上記第１の発明の話者特徴抽出装置と、上記話者特徴抽出装置によって抽出された周波数伸縮関数の逆関数を用いて標準話者の音声素片の周波数軸を伸縮する周波数ワープ手段で構成したので、話者の特徴をより的確に表す正確な周波数伸縮関数を用いて声質変換を行うことができる。したがって、発声データの内容に依存せずに安定して声質変換を行って、より入力話者の声質に近い合成音声を得ることができる。

【０１３７】また、第５の発明の話者特徴抽出方法は、所定の音声単位毎に周波数伸縮関数を最尤推定し、この推定された上記周波数伸縮関数の集合の頻度分布を求め、最大頻度を有する周波数伸縮関数を話者特徴として抽出するので、上記周波数伸縮関数の頻度分布に複数のピークが存在する場合でも、正確な周波数伸縮関数を安定して抽出することができる。したがって、この発明によれば、発声データの内容に依存せずに安定して話者特徴を抽出できる。

【０１３８】また、第６の発明のプログラム記録媒体は、コンピュータを、上記第１の発明における頻度計測手段およびモード抽出手段として機能させる話者特徴抽出処理プログラムが記録されているので、上記第１の発明の場合と同様に、上記周波数伸縮関数の頻度分布に複数のピークが存在する場合でも、正確な周波数伸縮関数を安定して抽出することができる。したがって、発声データの内容に依存せずに安定して話者特徴を抽出できる。

【図面の簡単な説明】

【図１】　この発明の話者正規化方式を用いた音声認識装置におけるブロック図である。

【図２】　図１における周波数ワープ関数推定部の学習時に機能する部分の詳細なブロック図である。

【図３】　図１における周波数ワープ関数推定部の認識時に機能する部分の詳細なブロック図である。

【図４】　周波数ワーピング関数の一例を示す図である。

30

【図５】　分布関数ｈ（α）の一例を示す図である。

【図６】　最大値を与えるαが複数個存在する分布関数ｈ（α）を示す図である。

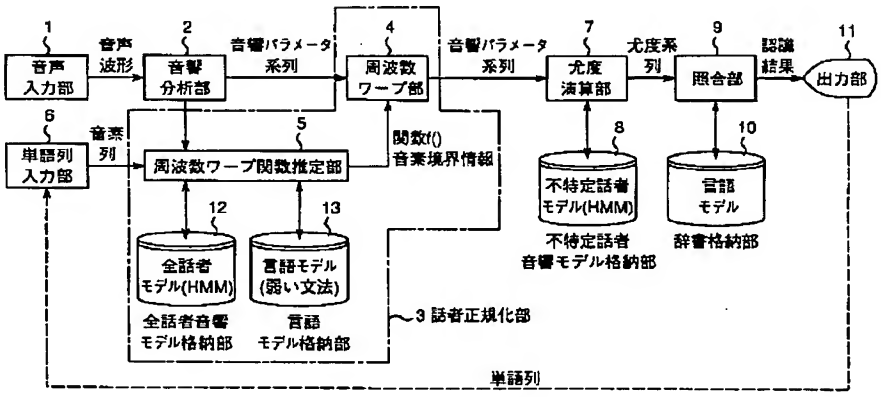
【図７】　図１とは異なる話者適応方式を用いた音声認識装置におけるブロック図である。

【図８】　この発明の音声合成装置のブロック図である。

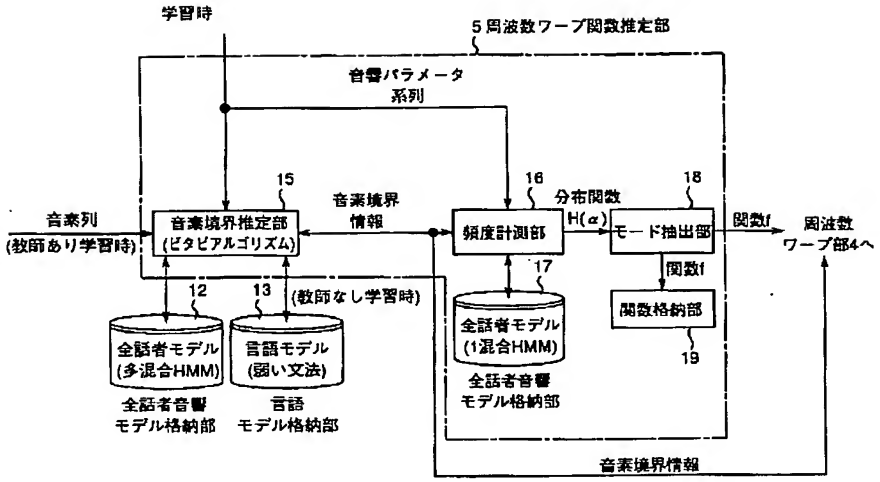
【符号の説明】

１，２１…音声入力部、
２，２２…音響分析部、
３…話者正規化部、
４，２５，４８…周波数ワープ部、
５，２４，４９…周波数ワープ関数推定部、
６，２６…単語列入力部、
７，２７…尤度演算部、
８，３４…不特定話者音響モデル格納部、
９，２９…照合部、
１０，３０…辞書格納部、
１１，３１…出力部、
１２，１７，３２…全話者音響モデル格納部、
１３，３３…言語モデル格納部、
１５…音素境界推定部、
１６…頻度計測部、
１８…モード抽出部、
１９…関数格納部、
２０…事前ワープ部、
２３…話者適応部、
２８…話者適応音響モデル格納部、
４１…テキスト解析部、
４３…韻律生成部、
４５…音声素片選択部、
４７…音声素片合成部、
５０…スピーカ。

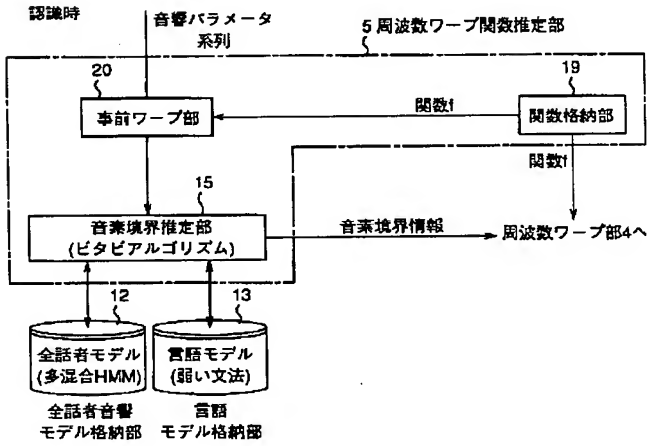
【図１】



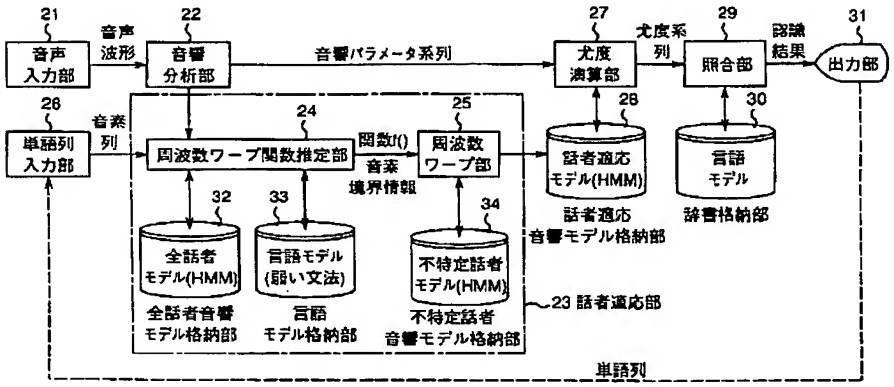
【図２】


学習時
5 周波数ワープ関数推定部
音響パラメータ系列
音素列
(教師あり学習時)
15
音素境界推定部
(ビタビアルゴリズム)
音素境界情報
16
頻度計測部
分布関数
H(α)
18
モード抽出部
関数f
周波数ワープ部4へ
関数f
関数格納部
19
17
全話者モデル
(1混合HMM)
全話者音響
モデル格納部
12
13
(教師なし学習時)
全話者モデル
(多混合HMM)
言語モデル
(弱い文法)
全話者音響
モデル格納部
言語
モデル格納部
音素境界情報

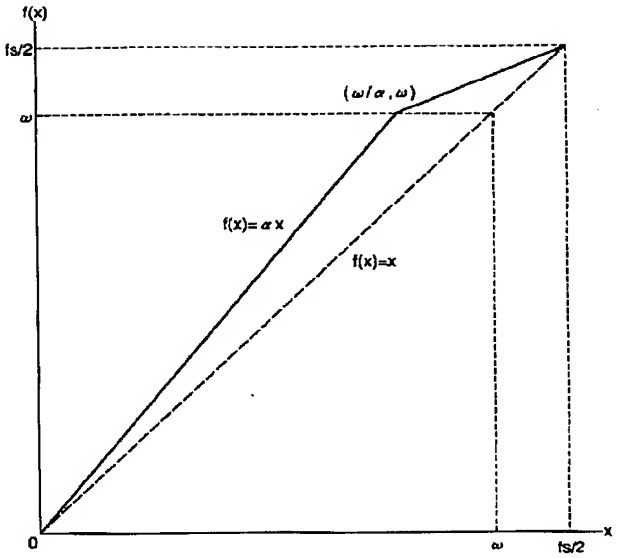
【図３】


認識時
音響パラメータ系列
5 周波数ワープ関数推定部
20
事前ワープ部
関数f
19
関数格納部
関数f
15
音素境界推定部
(ビタビアルゴリズム)
音素境界情報
周波数ワープ部4へ
12
13
全話者モデル
(多混合HMM)
言語モデル
(弱い文法)
全話者音響
モデル格納部
言語
モデル格納部

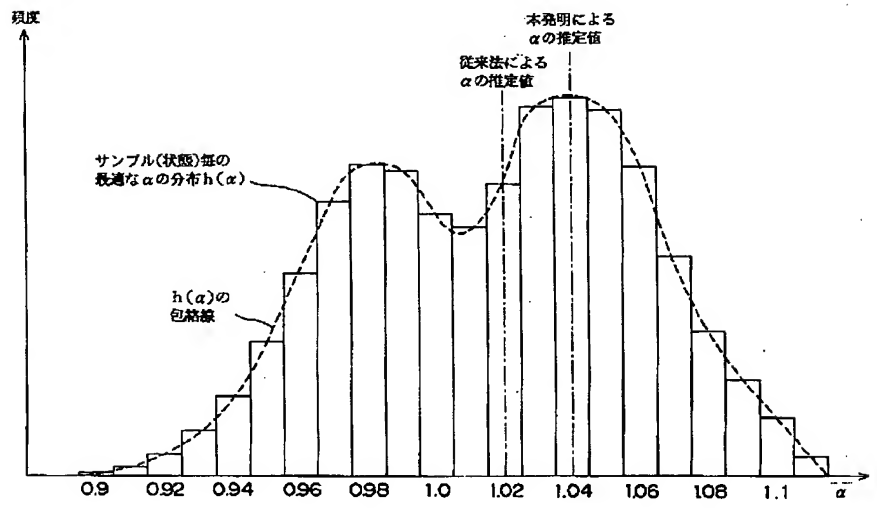
【図７】


21
音声入力部
音声波形
22
音響分析部
音響パラメータ系列
27
尤度演算部
尤度系列
29
照合部
認識結果
31
出力部
26
単語列入力部
音素列
24
周波数ワープ関数推定部
関数f()
音素境界情報
25
周波数ワープ部
28
話者適応モデル(HMM)
話者適応音響モデル格納部
30
言語モデル
辞書格納部
32
全話者モデル(HMM)
全話者音響モデル格納部
33
言語モデル(弱い文法)
言語モデル格納部
34
不特定話者モデル(HMM)
不特定話者音響モデル格納部
23 話者適応部
単語列

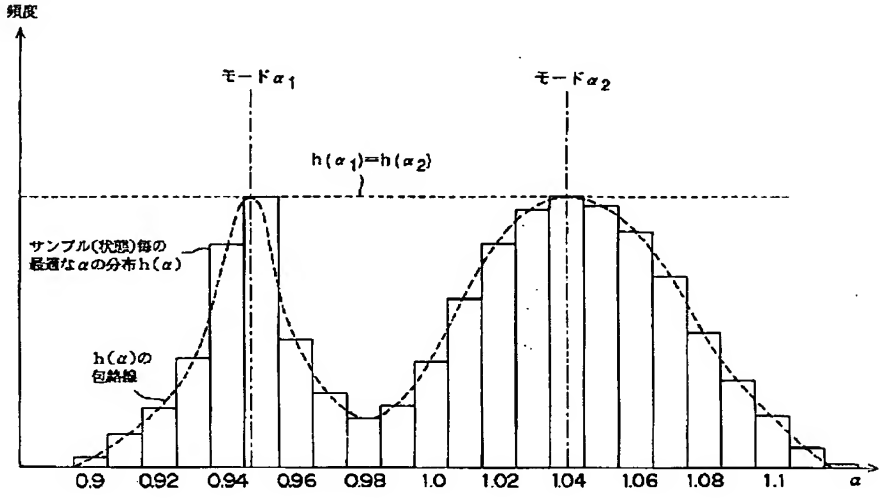
【図４】


f(x)
fs/2
ω
(ω/α, ω)
f(x)=αx
f(x)=x
0
ω
fs/2
x

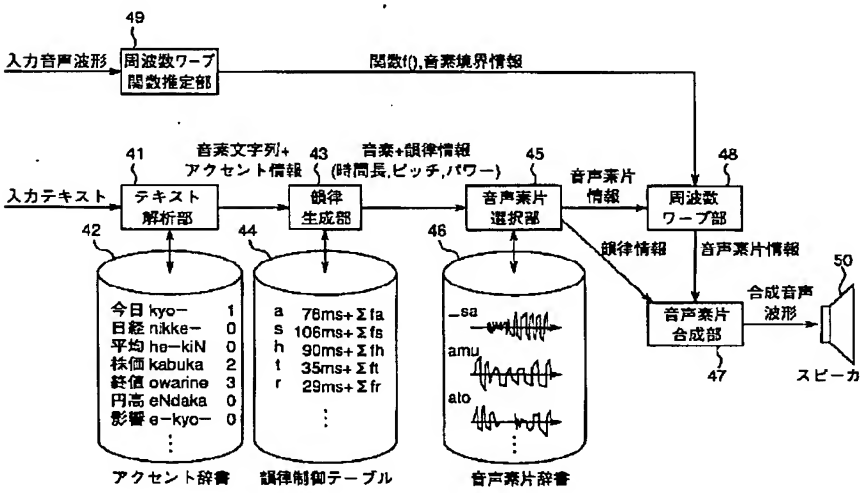
【図５】


頻度
本発明による
αの推定値
従来法による
αの推定値
サンプル(状態)毎の
最適なαの分布h(α)
h(α)の
包絡線
0.9
0.92
0.94
0.96
0.98
1.0
1.02
1.04
1.06
1.08
1.1
α

【図６】


頻度
モードα1
モードα2
h(α1)=h(α2)
サンプル(状態)毎の
最適なαの分布h(α)
h(α)の
包絡線
0.9
0.92
0.94
0.96
0.98
1.0
1.02
1.04
1.06
1.08
1.1
α

【図８】


49
入力音声波形
周波数ワープ
関数推定部
関数f(),音素境界情報
41
音素文字列+
アクセント情報
43
音素+韻律情報
(時間長,ピッチ,パワー)
45
音声素片
情報
48
入力テキスト
テキスト
解析部
韻律
生成部
音声素片
選択部
周波数
ワープ部
42
44
46
韻律情報
音声素片情報
50
合成音声
波形
今日 kyo− 1
日経 nikke− 0
平均 he−kiN 0
株価 kabuka 2
終値 owarine 3
円高 eNdaka 0
影響 e−kyo− 0
a 76ms+Σfa
s 106ms+Σfs
h 90ms+Σfh
t 35ms+Σft
r 29ms+Σfr
−sa
amu
ato
音声素片
合成部
47
スピーカ
アクセント辞書
韻律制御テーブル
音声素片辞書